Normalisation de concepts cliniques par des vecteurs de mots

François Morlane-Hondère[1] Cyril Grouin[1]
(1) LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay
{morlane,grouin}@limsi.fr

RÉSUMÉ

Dans cet article, nous présentons les expériences de normalisation de concepts cliniques (états pathologiques) que nous avons menées sur un corpus de messages postés sur des forums de santé. Le travail de normalisation consiste à identifier l'identifiant unique de concept (CUI) dans l'UMLS associé à chaque état pathologique présent dans le corpus. Nous avons réalisé cette normalisation au moyen de représentations vectorielles des mots présents dans le contexte des concepts cliniques (outil word2vec). Nous avons testé trois types de parcours des résultats produits par quatre modèles de voisins. Aucun type de parcours ne produit franchement de meilleurs résultats. En revanche, les modèles construits avec une fenêtre contextuelle de taille intermédiaire (entre 5 et 10 mots) permettent l'obtention des meilleurs candidats à la normalisation.

ABSTRACT

Clinical concept normalization using word embeddings

In this paper, we present the experiments we made to normalize clinical concepts (pathological states) found in messages from health forums. The normalization consists in identifying the concept unique identifier (CUI) from the UMLS to be associated with each clinical concept found in the corpus. We produced this normalization using a vectorial representation of words found in the neighborhood of clinical concepts (using word2vec). We tested three types of course in the lists of results produced by four models. We observed there is no type of course better than another one. Nevertheless, models built using a contextual window of 5 to 10 words (transitional size) allow to identify better candidates of normalization.

MOTS-CLÉS : Forums ; Normalisation ; Voisins distributionnels.

KEYWORDS: Forums ; Normalization ; Word Embeddings.

1 Introduction

Les effets indésirables dus aux traitements médicamenteux nécessitent une détection précoce en vue de les comprendre et de les prévenir. La pharmacovigilance est une activité de veille qui repose traditionnellement sur des déclarations réalisées par des patients auprès de professionnels de santé (médecins, pharmaciens, etc.), voire directement auprès d'un centre régional de pharmacovigilance. On estime que seules 4 à 5 % des déclarations sont réalisées de manière spontanée [1]. Les forums de santé, utilisés par les patients pour se renseigner, constituent une nouvelle source d'informations à exploiter pour la pharmacovigilance. Le style familier généralement adopté dans les messages postés sur les réseaux sociaux rend complexe l'identification des effets secondaires rapportés par les patients.

1. http://www.acadpharm.org/dos_public/GTNotif_Patients_Rap_VF__2015.01.22.pdf

Une fois ces effets identifiés par un système, ils doivent être évalués par les pharmacovigilants. Pour faciliter ce travail d'évaluation, il est nécessaire de normaliser ces expressions (*hypo → hypoglycémie*).

Dans cet article, nous abordons la question de la normalisation de concepts dans une terminologie médicale, l'UMLS (Lindberg *et al.*, 1993). Nous appliquons cette normalisation aux états pathologiques exprimés par les patients sur des forums de santé, en considérant que ces états pathologiques constituent des effets indésirables potentiels. Notre démarche se situe en aval d'un système d'identification des états pathologiques (Morlane-Hondère *et al.*, 2016a). L'objectif de cette étude vise à normaliser des expressions grand public au moyen de voisins distributionnels. Nous comparons ainsi trois types de parcours des résultats d'une recherche de meilleurs voisins, afin de normaliser au plus juste le maximum d'expressions d'états pathologiques.

La recherche de voisins distributionnels est une tâche permettant de regrouper des termes sur la base des contextes qu'ils partagent dans un corpus de textes. Cette tâche est traditionnellement utilisée pour la désambiguïsation sémantique (Gallant, 1991; Navigli, 2009), mais elle permet également la construction de ressources linguistiques (Claveau *et al.*, 2014; Ferret, 2015). Dans le domaine biomédical, elle est appliquée pour réduire la diversité des termes utilisés en corpus (Périnet & Hamon, 2014) et permet l'expansion d'abréviations (Wu *et al.*, 2015). Les approches actuellement utilisées en sémantique distributionnelle reposent sur le formalisme des réseaux neuronaux, notamment via l'implémentation *word2vec* (Mikolov *et al.*, 2013). C'est également ce type d'approche qui a été utilisée par Kaewphan *et al.* (2014) sur la tâche de normalisation de concepts cliniques dans l'UMLS lors de l'édition 2014 de SemEval.

2 Matériel et méthode

2.1 Extraction des entités PATHO

Le corpus dont nous disposons a été extrait par l'ISCOD (École des Mines de Saint-Étienne) dans le cadre du projet Vigi4MED. Il est constitué de 58,9 millions de messages postés sur des forums de santé entre juin 2000 et juin 2015, soit un total de 3,5 milliards de mots.

Dans le cadre de cette étude, nous avons travaillé sur un sous-ensemble du corpus ne comprenant que les messages dans lesquels a été identifié un nom de traitement lors d'une étape préalable d'identification des traitements et états pathologiques rapportés par les patients. Ce critère nous permet de nous assurer que les messages que nous prenons en compte traitent bien de problématiques de santé et ne relèvent pas de la discussion périphérique, ce qui risquerait de générer de l'ambiguïté lors du calcul des proximités distributionnelles. Ce filtrage est relativement restrictif puisque la taille du corpus s'en trouve réduite d'environ 74 %.

Dans cette étude, nous nous focalisons sur les entités qui correspondent à des états pathologiques (entités PATHO). Ces dernières ont été extraites à l'aide d'un système développé dans le cadre du projet. Le système, précédemment décrit dans Morlane-Hondère *et al.* (2016a), est basé sur la combinaison de deux approches statistiques :

— des champs conditionnels aléatoires (CRF) sont utilisés pour le repérage d'entités dites minimales : noms de pathologies (*cancer*, *rougeur*), parties du corps (*cerveau*, *bras*), etc. ;
— des machines à vecteurs de support (SVM) pour l'identification des cas où ces entités simples se combinent pour former des entités complexes.

Cette approche en deux étapes est connue dans la littérature sous le nom de *post-coordination* (Roberts *et al.*, 2015). Pour une description du système – notamment des traits utilisés –, se référer à Morlane-Hondère *et al.* (2016b).

La figure 1 résume le processus global que nous avons suivi pour extraire les entités (Morlane-Hondère *et al.*, 2016a), puis pour normaliser ces entités.

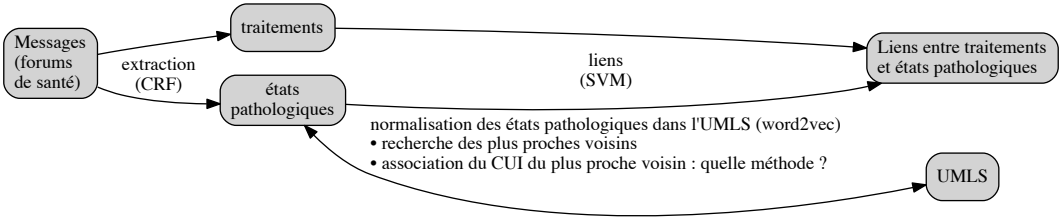

FIGURE 1 – Etape de normalisation des états pathologiques dans le processus global

2.2 Sélection des mots cibles

Le choix des mots cibles a été guidé par deux impératifs que sont la nécessité de choisir des mots dans des spectres de fréquence différents (les méthodes distributionnelles produisent de meilleurs résultats pour les mots les plus fréquents) et une contrainte de taille imposée par le fait que l'évaluation des voisins ne peut se faire que manuellement. De ce fait, nous avons scindé la liste des états pathologiques extraits par le système en trois groupes de fréquence – haute, moyenne et basse – et avons extrait aléatoirement, dans chacun de ces groupes, 40 états pathologiques absents de l'UMLS. Voici quelques exemples de mots cibles sélectionnés :

— fréquence haute : *stresse* (39 758), *crevée* (28 044), *hypo* (21 144), *accro* (20 937) ;
— fréquence moyenne : *saigne* (11 953), *brule* (6003), *croutes* (2479), *mal au bas ventre* (2000) ;
— fréquence basse : *rhino pharyngite* (332), *doigts enflés* (53), *manque de sucre* (49), *crampes atroces* (46).

L'absence de ces états pathologiques dans l'UMLS peut s'expliquer par le fait qu'elles correspondent à des formes mal orthographiées, à des expressions non standard (apocopes, expressions imagées ou paraphrase de concepts médicaux en langue courante), ou au fait qu'il s'agisse de formes fléchies de concepts apparaissant dans l'UMLS.

Comme on pouvait s'y attendre, la pertinence des entités extraites décroît en même temps que leur fréquence d'apparition dans le corpus. Ainsi, parmi les mots cibles de basse fréquence figurent plusieurs entités qui sont des erreurs d'extraction comme *dos crawlé* (incorrectement extrait sur le modèle d'une combinaison ANATOMIE + PATHOLOGIE comme *dos bloqué* ou *dos coincé*) ou *negus* (dont la terminaison est semblable à la terminaison latine *-us* récurrente dans les pathologies, commme dans *infarctus* ou *prolapsus*). Bien qu'aucune normalisation ne soit possible pour ces entités, nous avons choisi de les conserver dans notre jeu d'évaluation en considérant la démarche de normalisation comme directement tributaire des résultats générés en amont par un système d'extraction d'entités dont il faut assumer les imperfections.

2.3 Génération et annotation des voisins distributionnels

Un vecteur de mots a été généré à l'aide de word2vec pour chacune des 44 865 entités PATHO dont la fréquence est supérieure à 20. La taille des vecteurs a été fixée à 200. La taille de la fenêtre de mots est un critère extrêment important dans la génération d'un modèle distributionnel. Des études ont ainsi montré qu'une fenêtre restreinte avait tendance à extraire des relations relevant de la similarité sémantique alors qu'une fenêtre large permettait de ramener des relations associatives plus lâches. Afin de vérifier quel type de modèle est le plus adapté à notre tâche de normalisation, nous avons généré quatre jeux de vecteurs avec des tailles de fenêtre de 1, 5, 10 et 20 (modèles respectivement nommée par la suite *w1*, *w5*, *w10* et *w20*).

Une fois, les vecteurs générés, nous avons à nouveau utilisé word2vec pour calculer la similarité cosinus entre les vecteurs et ainsi extraire les 20 meilleurs voisins distributionnels pour chacun des états pathologiques.

Les voisins ont ensuite subi une double annotation. Dans un premier temps, nous avons identifié automatiquement ceux qui apparaissent dans notre ressource de référence, l'UMLS. Ces voisins constituent des normalisations potentielles. Nous avons ensuite distingué manuellement les normalisations valides des normalisations fautives (*prise de poids* vs *anorexique* pour *grossir*).

2.4 Heuristiques de parcours des voisins

Après avoir généré et classé par similarité décroissante les voisins distributionnels de chacune des entités à normaliser, il est possible de parcourir ces listes de plusieurs manières. Nous avons envisagé trois types de parcours, avec pour objectif d'optimiser l'accès à la meilleure normalisation :

— le parcours en profondeur consiste à parcourir les n premiers voisins d'une entité à normaliser, et à afficher ces voisins :

```
for entité do
    forall the 20 premiers voisins do
        afficher les voisins
    end
end
```

— le parcours en largeur consiste à parcourir, pour chacun des n premiers voisins, les m premiers voisins qui lui sont associés (les voisins des voisins) et à les afficher, puis à réitérer pour chaque voisin de premier niveau :

```
for entité do
    forall the 4 premiers voisins do
        afficher le voisin;
        forall the 4 premiers voisins du voisin do
            afficher les voisins
        end
    end
end
```

— le parcours mixte est une variante du précédent et consiste à d'abord parcourir les n premiers voisins, avant de parcourir les m premiers voisins de chacun des voisins de premier niveau :

```
for entité do
    forall the 4 premiers voisins do
        afficher les voisins
    end
    forall the 4 premiers voisins do
        forall the 4 premiers voisins du voisin do
            afficher les voisins
        end
    end
end
```

Chacune de ces heuristiques correspond à une stratégie différente. Alors que le parcours en profondeur consiste à parcourir les n meilleurs voisins d'une entité par score de similarité décroissant, le parcours en largeur repose sur l'idée selon laquelle il est possible d'accéder à des formes normalisées de façon plus efficace via des voisins de voisins – voisins *de deuxième degré* – de haut rang plutôt qu'en parcourant l'intégralité des voisins de premier degré de l'entité à normaliser. Comme son nom l'indique, le parcours mixte combine ces deux approches en parcourant dans un premier temps les quatre meilleurs voisins de premier degré du mot cible avant de consulter leurs meilleurs voisins. Ainsi, les voisins examinés par le parcours en largeur et le parcours mixte sont les mêmes, la différence résidant dans l'ordre de parcours.

La liste des voisins parcourus par chacune de ces règles pour le mot cible *pertes d'audition* (modèle w1) est rapportée au tableau 1. Les voisins qui apparaissent en gras sont ceux qui apparaissent dans l'UMLS. Ils correspondent donc à des normalisations potentielles. Les voisins en gras soulignés sont ceux que nous avons considérés comme des normalisations valides. On constate que, du fait de la différence d'ordre de parcours des voisins entre les parcours en largeur et mixte, *pertes auditives* – que nous avons considéré comme une normalisation valide – est parcouru en quatrième dans le modèle mixte alors qu'il est relégué en seizième position dans le modèle en largeur. On remarque également que le parcours des voisins de deuxième degré entraîne la présence de doublons dans les voisins rapportés.

3 Résultats et discussion

Le parcours en profondeur des résultats est le plus intuitif : il consiste à parcourir les n premiers voisins associés à l'entité à normaliser. C'est généralement ce type de parcours qui apporte les meilleurs résultats (voir tableau 1, deux concepts jugés pertinents par les humains).

Le parcours en largeur cherche à maximiser les meilleurs voisins, non seulement en se limitant aux quatre premiers voisins d'une entité à normaliser, mais également en étudiant les quatre premiers voisins de chacun de ces voisins de haut niveau. Ce type de parcours, parce qu'il s'intéresse aux voisins positionnés dans les premiers rangs, permet de retrouver le plus de candidats potentiels (cf. 9 voisins présents dans l'UMLS dans le parcours en profondeur contre 11 voisins dans l'UMLS dans le parcours en profondeur). En revanche, ce mode de parcours est susceptible de ramener plusieurs exemplaires du même candidat, si ce candidat est un voisin de plusieurs voisins (par exemple, *hyperacousie* dans le tableau 1).

Enfin, le parcours mixte reprend les mêmes voisins que ceux retenus dans le parcours en largeur (le nombre de voisins présents dans l'UMLS est donc le même). La différence de ce type de parcours

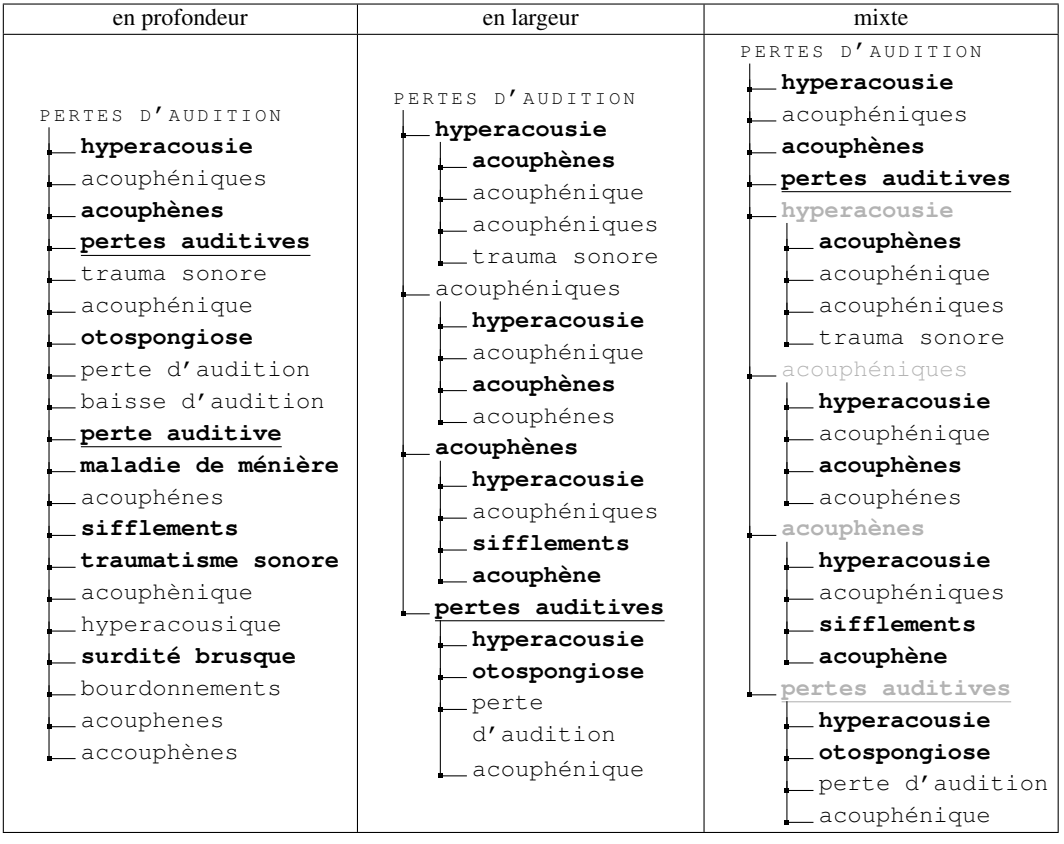

TABLE 1 – Voisins ramenés par les trois règles pour le mot cible *pertes d'audition* dans le modèle w1. En gras, les normalisations candidates (entités présentes dans l'UMLS). Les normalisations que nous avons jugées valides ont été soulignées

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	total
w1	ver	6	1	3	5	3	0	4	4	0	2	2	1	2	0	0	2	0	1	0	0	36
	hor	6	2	3	5	5	0	0	0	3	1	1	1	0	0	1	2	0	0	0	0	30
	mix	6	1	3	5	2	2	3	2	0	0	3	1	1	0	0	1	0	0	0	0	30
w5	ver	5	7	5	3	8	0	2	1	3	0	0	0	0	2	0	1	0	0	1	1	39
	hor	5	1	3	2	4	2	3	2	2	1	2	0	0	0	0	1	2	2	0	0	32
	mix	5	7	5	3	0	0	0	3	2	0	2	1	0	0	0	0	2	2	0	0	32
w10	ver	5	4	4	5	4	5	1	0	1	1	1	0	0	1	1	0	0	0	0	0	33
	hor	5	0	2	2	9	1	0	0	1	0	2	1	1	1	0	4	0	0	0	2	31
	mix	5	4	4	5	0	1	0	7	0	0	1	0	0	1	1	0	0	0	0	2	31
w20	ver	4	3	6	3	0	2	1	4	1	3	1	2	0	0	2	3	0	0	0	1	36
	hor	4	2	1	7	2	1	0	2	1	4	0	0	1	0	0	1	0	0	0	0	26
	mix	4	3	6	3	1	1	2	1	0	2	0	2	0	1	0	0	0	0	0	0	26

TABLE 2 – Rang des normalisations candidates en fonction du modèle utilisé (w1...w20) et du type de parcours (en profondeur, en largeur, mixte) des listes de résultats

repose sur l'ordre dans lequel sont présentés les voisins : d'abord les quatre premiers voisins verticaux, puis les quatre premiers voisins de ces voisins de haut niveau. L'intérêt de ce type de parcours vise à d'abord présenter les meilleurs voisins les plus proches de l'entité à normaliser avant de s'intéresser aux voisins de ces meilleurs voisins.

Le tableau 2 indique, pour chaque modèle distributionnel et chaque règle, les rangs auxquels apparaissent les normalisations valides. On peut ainsi voir que, pour le modèle w1 et le parcours en profondeur, 6 normalisations valides sont des voisins de rang 1. On constate que, globalement, les voisins qui correspondent à des normalisations valides se retrouvent dans la partie haute du classement : en moyenne, sur l'ensemble des modèles, 60 % des normalisations valides d'un mot cible donné se retrouvent parmi ses cinq meilleurs voisins, et 82 % parmi ses dix meilleurs voisins. Cela tend à montrer que, dans le cadre d'une normalisation par proximité distributionnelle, quand une forme possède une normalisation parmi ses voisins, le haut degré de proximité distributionnelle entre les deux fait qu'il devient peu pertinent de chercher des formes normalisées au-delà d'un certain seuil (que l'on peut fixer autour de 10). Ainsi, du fait que les normalisations valides figurent la plupart du temps parmi les cinq plus proches voisins des mots cibles, on ne s'attend pas à des différences spectaculaires entre les trois règles que nous proposons.

Le tableau 3 indique, pour chaque règle et chaque modèle, le nombre de cas où la première normalisation rencontrée est valide. Ces chiffres permettent de mesurer l'intérêt d'une stratégie qui consisterait à associer systématiquement chaque entité à normaliser avec le premier de ses voisins qui possède un CUI. Les résultats montrent des différences marquées entre les différents modèles distributionnels. Les modèles dont les fenêtres sont la plus étroite – w1 – ou la plus large – w20 – sont ceux qui apparaissent les moins adaptés à cette stratégie. Les modèles intermédiaires – w5 et w10 – obtiennent des performances légèrement supérieures puisque leur utilisation permettrait de normaliser automatiquement quasiment 20 mots cibles – soit environ 17 % du jeu d'exemples. Comme prédit dans le paragraphe précédent, la différence entre les trois règles n'apparaît pas ici de façon marquée.

règle	w1	w5	w10	w20	moyenne
en profondeur	15	22	19	15	17,75
en largeur	17	17	20	14	17
mixte	16	20	20	16	18
moyenne	16	19,7	19,7	15	

TABLE 3 – Nombre de cas où la première normalisation rencontrée est valide

4 Conclusion

Dans cet article, nous avons présenté les expériences que nous avons menées pour normaliser des expressions d'états pathologiques exprimées par des patients sur des forums de santé. Notre approche prend en entrée le résultat d'une identification de traitements et d'états pathologiques (ce travail a été décrit dans un précédent article). L'approche de normalisation que nous avons envisagée repose sur l'étude de voisins distributionnels, calculés pour chaque entité à normaliser, au moyen de l'outil word2vec. Nous avons testé quatre modèles de voisins distributionnels en faisant varier la taille de la fenêtre autour des entités à regrouper (des tailles de 1, 5, 10 et 20 mots).

Afin d'optimiser l'identification des voisins candidats pour la normalisation des entités, nous avons

comparé trois modes de parcours des résultats de voisins : (i) un mode en profondeur reposant sur les n premiers voisins classés par similarité décroissante, (ii) un mode en largeur reposant sur les m premiers voisins de chacun des n premiers voisins (permettant de se focaliser sur les meilleurs voisins de premier et deuxième niveau), et (iii) un mode mixte, consistant à d'abord s'intéresser aux n premiers voisins (de premier niveau) avant de prendre en compte les m premiers voisins (de deuxième niveau) de chacun des voisins de premier niveau. Le résultat de nos expériences ne laisse pas entrevoir un type de parcours franchement meilleur qu'un autre. Il apparait que le mode en profondeur (le plus simple) permet d'identifier un peu moins de candidats, mais ceux identifiés sont pertinents en regard de l'expression à normaliser.

Nous observons en revanche que, parmi les différents modèles testés, les modèles avec une taille de fenêtre trop restreinte (1 seul mot) ou trop large (20 mots) sont parmi les moins performants dans l'obtention de candidats à la normalisation. En effet, nous obtenons nos meilleurs résultats avec les modèles construits à partir de fenêtres de 5 et 10 mots, sans qu'il ne fut possible d'identifier un modèle meilleur qu'un autre parmi ces deux modèles.

Remerciements

Ce travail a été réalisé dans le cadre du projet Vigi4MED (financement Agence Nationale de Sécurité du Médicament sous le numéro ANSM-2013-S-060).

References

CLAVEAU V., KIJAK E. & FERRET O. (2014). Explorer le graphe de voisinage pour améliorer les thésaurus distributionnels. In *Actes TALN*, Marseille, France.

FERRET O. (2015). Déclasser les voisins non sémantiques pour améliorer les thésaurus distributionnels. In *Actes TALN*, Caen, France.

GALLANT S. I. (1991). A practical approach for representing context and for performing word sense disambiguation using neural networks. *Neural Computation*, **3**(3), 293–309.

KAEWPHAN S., HAKAKA K. & GINTER F. (2014). UTU: disease mention recognition and normalization with CRFs and vector space representations. In *Proc of SemEval*, p. 807–11, Dublin, Ireland.

LINDBERG D. A., HUMPHREYS B. L. & MCRAY A. T. (1993). The Unified Medical Language System. *Methods Inf Med*, **32**(4), 281–91.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. & DEAN J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances on Neural Information Processing Systems*.

MORLANE-HONDÈRE F., GROUIN C. & ZWEIGENBAUM P. (2016a). Identification of adverse drug reactions from social media. In *Proc of LREC*, Portorož, Slovenia.

MORLANE-HONDÈRE F., GROUIN C. & ZWEIGENBAUM P. (2016b). Représentation des informations textuelles pour la détection d'états pathologiques par apprentissage statistique. In *Actes des JFIM*, Genève.

NAVIGLI R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys*, **41**(2).

PÉRINET A. & HAMON T. (2014). Réduction de la dispersion des données par généralisation des contextes distributionnels : application aux textes de spécialité. In *Actes TALN*, Marseille, France.

ROBERTS K., RODRIGUEZ L. M., SHOOSHAN S. E. & DEMNER-FUSHMAN D. (2015). Automatic extraction and post-coordination of spatial relations in consumer language. In *Proc of AMIA*, p. 1083–1092, Washington, DC.

WU Y., XU J., ZHANG Y. & XU H. (2015). Clinical abbreviation disambiguation using neural word embeddings. In *Proc of BioNLP*, p. 171–6, Beijing, China.